# Normalization of RNA sequence data

Normalization is a process of making data comparable within or across samples by minimizing the technical variations and biases and retaining the biological variation alone (Bullard et al. 2010; Risso et al. 2014; Robinson & Oshlack 2010; Zhao, Yingdong et al. 2021; Zwiener, Frisch & Binder 2014). It is one of the crucial steps which impacts the results of data analysis (Bullard et al. 2010; Evans, Hardin & Stoebel 2017; Risso et al. 2011; Robinson & Oshlack 2010; Zwiener, Frisch & Binder 2014). Se. (other important steps are statistical testing of hypothesis and alignment which will be mentioned in other documents). Several benchmarking studies on normalization have been published however, the consensus which is the most accurate is lagging (Bullard et al. 2010; Dillies et al. 2013; Li et al. 2015; Lin et al. 2016; Maza et al. 2013; Zwiener, Frisch & Binder 2014; Zyprych-Walczak et al. 2015). The normalization techniques can be assessed by comparing the results with real-time qPCR for benchmarking the true expression values (Li et al. 2015). Assessing the suitability of a selected normalization method can be challenging, but one approach is to employ multiple methods and evaluate the consistency of the outcomes depending on the research question.

# Intrasample normalization techniques

Intrasample normalization is the comparison between the expression of different genes within a sample. In such a comparison, the inherent features of a gene such as effective gene length and GC content needs to be accounted. Effective gene length is the length of gene excluding its intron region. Genes with shorter effective length have lower counts and higher rate of dropouts (genes are expressed but not detected) (Oshlack & Wakefield 2009; Phipson, Zappia & Oshlack 2017). RPKM/FPKM (Reads/Fragments per kilo base of transcript per Million mapped reads) and TPM (Transcripts per million) are widely used techniques for intrasample comparison. RPKM is an expression unit first described by (Mortazavi et al. 2008) for single-end reads whereas FPKM is used for paired-end reads (Trapnell et al. 2010). RPKM corrects library size first and then gene length variance. However, evidence showed that while attempting to adjust for variations in gene length during a differential analysis, it led to a biased impact on the variances per gene, especially for genes with low expression levels (Oshlack & Wakefield 2009).In addition, it also produces inconsistent results among the samples (Wagner, Kin & Lynch 2012). Unlike RPKM, the TPM, another method for intrasample comparison, corrects gene length followed by the library size. Hence, the resulting expression values reflects the true “relative molar concentration” (Wagner, Kin & Lynch 2012). The assumption of RPKM and TPM is that production and distribution of mRNAs across the samples is same. However, with the change in condition, amount of mRNA production also changes. This is main reason why RPKM and TPM can be used to compare genes only within sample or replicates of sample where the assumption of same amount and distribution of mRNA holds true (Zhao, S, Ye & Stanton 2020).

RPKM/FPKM =

Here, 103 is for gene length and 106 is for sequencing depth factor.

TPM=A∗ ∗

Where, A =

Here, 106 is for sequencing depth factor.

Similar to length causing the unwanted variation in count data, GC is another factor, that inherently cause the variation between different genes. It has been found that the genes with higher or lower GC content are underrepresented during sequencing ((Benjamini & Speed 2012; Risso et al. 2011). A GC content bias is observed partially due to PCR amplification while preparing library (Aird et al. 2011). Aother reason is genome functionality (Risso et al. 2011). GC content is confounding effect which is nonlinear and highly sample specific (Love, Hogenesch & Irizarry 2016). To account for GC bias, a within sample gene level approach is proposed in *EDASeq* package, three methods are implemented here such as regression normalization, global scaling normalization and quantile normalization and these techniques mainly stratify genes in same sized bins based on GC content, then match the count distributions across bins (Risso et al. 2011). GC effect is not consistent between repeated experiments making it difficult to account for (Benjamini & Speed 2012).

# Intersample normalization techniques

Intersample comparison is comparison of a gene present in two different samples. This type of comparison should correct sequencing depth variance and composition bias (Bullard et al. 2010; Robinson & Oshlack 2010), batch effects along with other known and unknown technical factors needs to be corrected. Sequencing depth, also known as, library size is total read counts mapped for a given sample. There is no optimal limit for sequencing depth as it depends on complexity of the organism and aim of the experiment (Conesa et al. 2016). A deep sequencing might improve the quantification and identification of transcripts leading to novel isoform detection; however, it may also detect transcriptional noise and off-target transcripts (Conesa et al. 2016). Sequencing different samples within a single experimental design may have different sequencing depths because total number of representative RNA molecules in cDNA library can vary (Dillies et al. 2013).If RNA sample has a small number of highly expressed genes, the reads representing those genes will take away substantial resources from the sample, which decreases the reads for genes causing under sampling, if not explicitly accounted for, this can cause the composition bias to the detection of differentially expressed features (Robinson & Oshlack 2010).

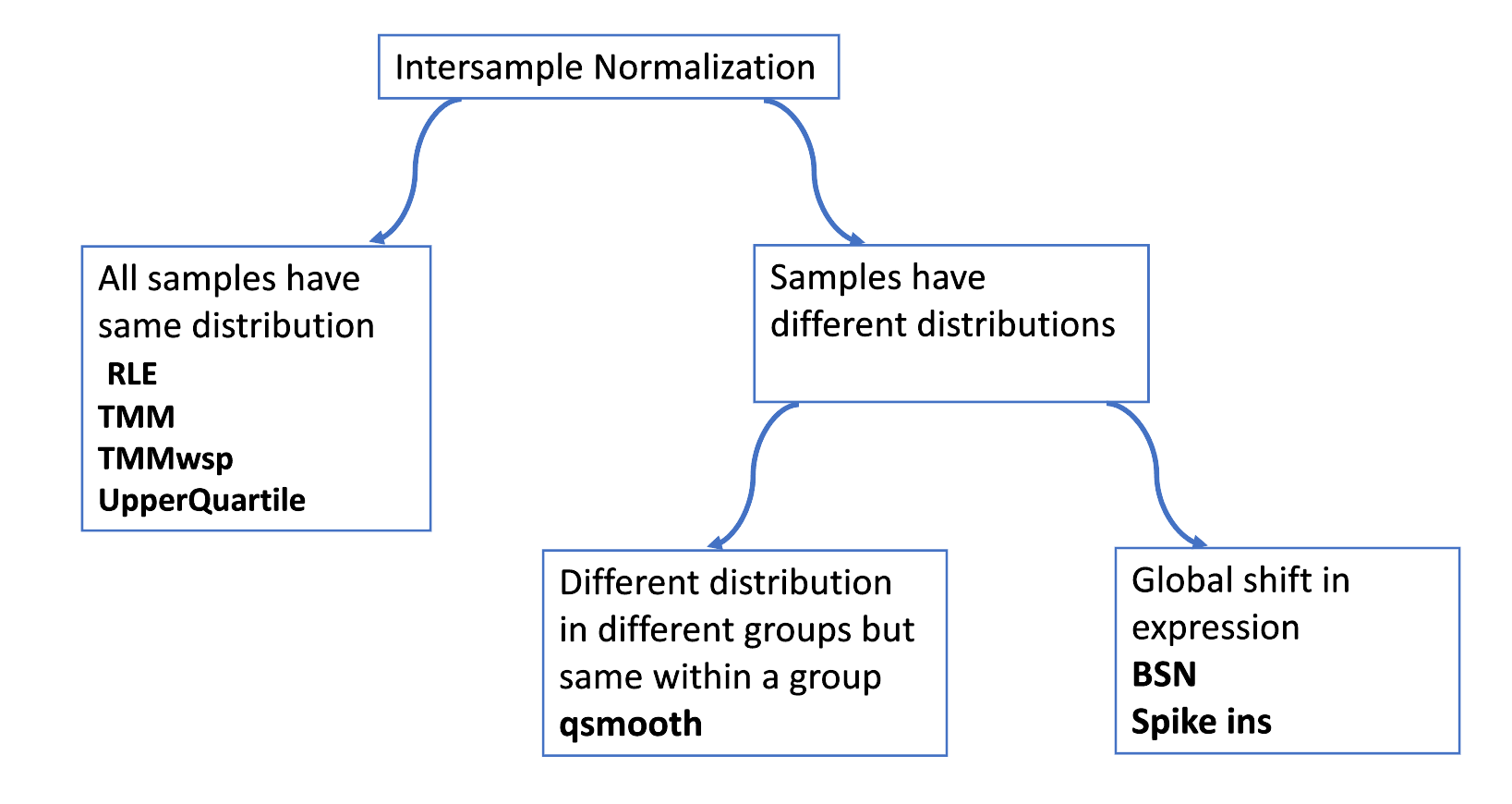


Fig 1: Techniques for intersample normalization based on distribution of samples.

TMM (trimmed mean of M-values), RLE (relative log expression) and quantile normalization are widely used tools for intersample comparison which assumes all the samples have similar distribution and this assumption holds true when majority of genes are not dysregulated (Evans, Hardin & Stoebel 2017; Zhao, Yaxing, Wong & Goh 2020). TMM, RLE, and quantile normalization are global adjustment methods which force all the observed distribution of all the samples to be same and using these methods without understanding the actual distribution of all samples will distort the true biological signal (Evans, Hardin & Stoebel 2017; Wang et al. 2011; Zhao, Yaxing, Wong & Goh 2020). TMM (Robinson & Oshlack 2010), which is incorporated into the *edgeR* package, is widely used for differential gene expression. CPM (Counts per million mapped reads) is a basic gene expression unit (Chen, Lun & Smyth 2016; Law et al. 2018). CPM corrects the sequencing depth variation.

CPM =

The TMM method further helps to correct the compositional bias, a reference sample is chosen as a basis for comparison. *edgeR* chooses the sample as a reference sample which is near to the average mean of all the samples. Then, genes are selected to create a scaling factor, for this, the genes which are not outlying are selected (outliers can be known by observing log fold differences). Log fold values are trimmed from top 30% and bottom 30% and mean of logs are trimmed from top 5% and bottom 5%. The genes which are overlapping from both trimmed log fold values and mean of logs are used to calculated scaling factor. Finally, the original read counts are divided by this scaling factor. How TMM is chosing a reference sample and trimming its count values to create a scaling factor is the major difference with other techniques and this makes it as robust technique to handle outliers. (Robinson & Oshlack 2010). TMMwsp (TMM with singleton pairing) is a variant of TMM to handle the zero inflated data. In TMM, gene having zero in either library while comparing a pair of libraries is ignored, but in TMMwsp it is not ignored. Therefore, TMMwsp is suitable for data with inflated zero counts. The RLE in DESeq2 Bioconductor package is similar to the TMM, in which, first a **reference sample is created by calculating row-wise geometric mean which can handle outlier data. Subsequently, scaling factor is calculated by using the median of the ratio for each sample in log base. The median is another way to handle extreme values. Then, median is converted to normal numbers from log base to get the scaling factor**. Finally, count values are normalized using these scaling factor (Love, Huber & Anders 2014). Quantile normalization was first proposed for microarray data. Quantile normalization is achieved by replacing each quantile in the data with the average or median value of that quantile, considering the values from all samples (Bolstad et al. 2003). This process aims to align the distributions and make them comparable across different samples. There are two types of quantile normalization, the upper quartile (UQ) normalization which is determined by applying the 0.75 quantile to the gene counts from all runs and is applied in sequencing data. Scale factors are calculated from the 75% quantile of the counts of each library (Bullard et al. 2010).

In large scale studies and studies involving different cells from different organs or cancer cells, the assumptions of TMM, RLE and Quantile methods might fail, where large number of genes might be dysregulated, in this type of data, all the samples may not have similar distributions and majority of genes might be upregulated in one biological condition relative to other condition. For the normalization of global shift in expression and unbalanced number of upregulated and downregulated genes, either reference genes whose expression is constant, or spike ins can be used (Liu et al. 2019; Lovén et al. 2012). Another method for global shift in expression is BSN (Biological scaling normalization) which needs experimentally measured polyA+ RNA amounts . In some cases, distribution of samples with a group is same but across all the groups is different. For this situation, a normalization technique has been developed called smooth quantile normalization or *qsmooth* (Hicks et al. 2017).

## 3. Batch effects and unwanted variations normalization

Although increase in sample size gives more statistical power to analysis, the chances of getting batch effects and other unwanted variations also increases proportionally. There might be known and unknown batch effects, according to their nature, they can be accounted and corrected. There are mainly two ways, one way is correcting the batch effects first and separately modelling the corrected data with covariates of interest and the other way is accounting for the unknown batch effects during data modelling of both covariates of interest and batch variations together. The batch effect appears due to difference on experimental design, different sequencing platforms, and researcher which causes unwanted noise in data. (Bullard et al. 2010; Leek et al. 2010).

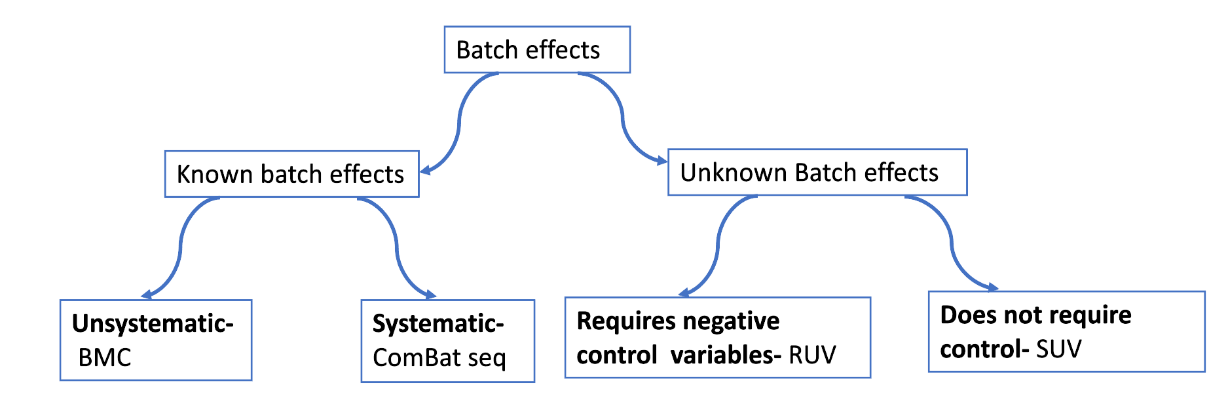


Fig 2: Techniques for accounting and correcting batch effects based on nature of unwanted variation.

BMC (Batch mean cantering) and Combat seq are techniques for correcting the known batch effects. These are practical and have broad applications. ComBat seq is implemented in *SVAseq* package.

RUV (remove unwanted variation) and SVA (surrogate variable analysis) seq find the latent factors due to batch effect and use it as a covariate in the data modelling. RUV was developed to identify and adjust uncharacterized technical variations in data analysis using negative control variables. RUVseq follows generalised linear model (GLM) where both biological covariates and unknown factors due unwanted variation are regressed. Three approaches are used to estimate the unknown factors, they are, RUVg (negative control genes are used which have constant expression across samples), RUVs (negative control sample/replicate which is assumed to have covariates of interest are constant) and RUVr (residuals which are from first pass GLM regression on the known covariates). However, the use of spike ins as negative control is difficult, since it is challenging to maintain constant amount of spike ins in all samples in same ratio (Robinson & Oshlack 2010) and it is still unclear if the synthetic spike ins act in same manner as cellular transcripts (Grün & van Oudenaarden 2015). SVAseq builds a surrogate variable that can be used to adjust unknown variables due to technical noise. It is univariate method and based on Gaussian likelihood assumption. Data is modelled as the combination of known covariates of interest, known adjustment variables and unknown artifacts. Here, first a subset of data affected by artifact is identified and then matrix decomposition like is applied to estimate the artifacts, the resulted estimate is used to correct the analysis (Leek 2014). The genomic data affected by artefact is determined by identifying the variables which are not associated with biological variables if interest. This can be done by fitting a null model and full linear model to the data.

**Comparison of different types of normalization methods for Differential gene expression**

# Methods

We are particularly interested to compare the methods for differential gene expression. From above literature survey we know that TMM, RLE and UQ are major normalization techniques for differential gene expression analysis. We are adding a variant of TMM normalization technique called TMMwsp.

**Dataset**

All normalization strategies were compared to each other by implementing on the same data set. The data was obtained from a paper “Generation of a microglial developmental index in mice and in humans reveals a sex difference in maturation and immune reactivity” published in Glia in 2017 and subsequent erratum in 2018 (Hanamsagar et al. 2018). The data was downloaded from <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99622>.  In this analysis, gene level differential analysis was performed to look the developmental changes only in Microglial cells, and immune activated samples as well. The total sample size of the study is 60, however, we are sub-setting the sample to study the developmental phase and immune activated phase separately.

For the developmental phase: E18, P4, P14 and P60\_sal with total 47 samples

For LPS and Saline treated cells: P60\_sal and P60\_LPS with total 25 samples

**Pre-processing:**

1. **Loading and Filtering of lowly expressed genes**

The raw data was retrieved from NCBI and subsequently saved in a directory. The raw data was made read only so that it would not be manipulated during the data munging process.

# Make data immutable from terminal

% chmod 400 GSE99622\_hanamsagar2017\_raw\_reads.txt

The data was loaded with genes as row names and samples as column names. edgR function *filterByExpr* was used to filter lowly expressed genes*.* This function keeps the rows with worthwhile counts. This filtering keeps genes that have CPM above “k” in “n” samples, here “k” minimum count according to the library size and “n” is the smallest group size (Chen, Lun & Smyth 2016).

keep <- filterByExpr (y, design)

y <- y[keep,]

1. **Choosing negative control**

Negative control is CPM normalized counts which is log transformed. All other methods of normalization were used after doing CPM normalization which adjust the sequencing depth bias and are log transformed. Method is “none” for no normalization that gave negative control. “TMM” for TMM normalization, “TMMwsp” for TMMwsp normalization, “RLE” for RLE normalization and “upperquartile” for UQ normalization.

no\_normalise <- calcNormFactors(y, method = "none")

logcount <- cpm(no\_normalise,log=TRUE)

1. **Exploratory data analysis**

The histogram and box plots of all the samples using all four methods and negative control were created to know the distribution of the sample.

1. **Getting the differentially expressed genes**

The differentially expressed genes (DEGs) using all four methods and negative control were obtained using *edgR*. Here for example, DEGs from UQ is being calculated.

UQ <-estimateDisp(UQ,design, robust = TRUE)

UQ <-glmQLFit(UQ,design, robust =TRUE)

UQ <-glmQLFTest(UQ, contrast = E18vsP4)

The *estimateDisp* function maximizes the negative binomial likelihood to give the estimation of common, trended, gene wise dispersion across all the genes, then empirical Bayes method was used to estimate posterior dispersion, robust = TRUE means, the estimate should be robust against the outliers. *glmQLFFit* function performs empirical Bayes quasi likelihood F-Tests. The contrast matrix was set E18 – P4, as an example, other matrix of interest can also be chosen. Multiple test correction was performed by applying the Benjamini-Hochberg method on the P values to control False positive discovery rate (FDR).

1. **Testing the Batch effect correction on the data**

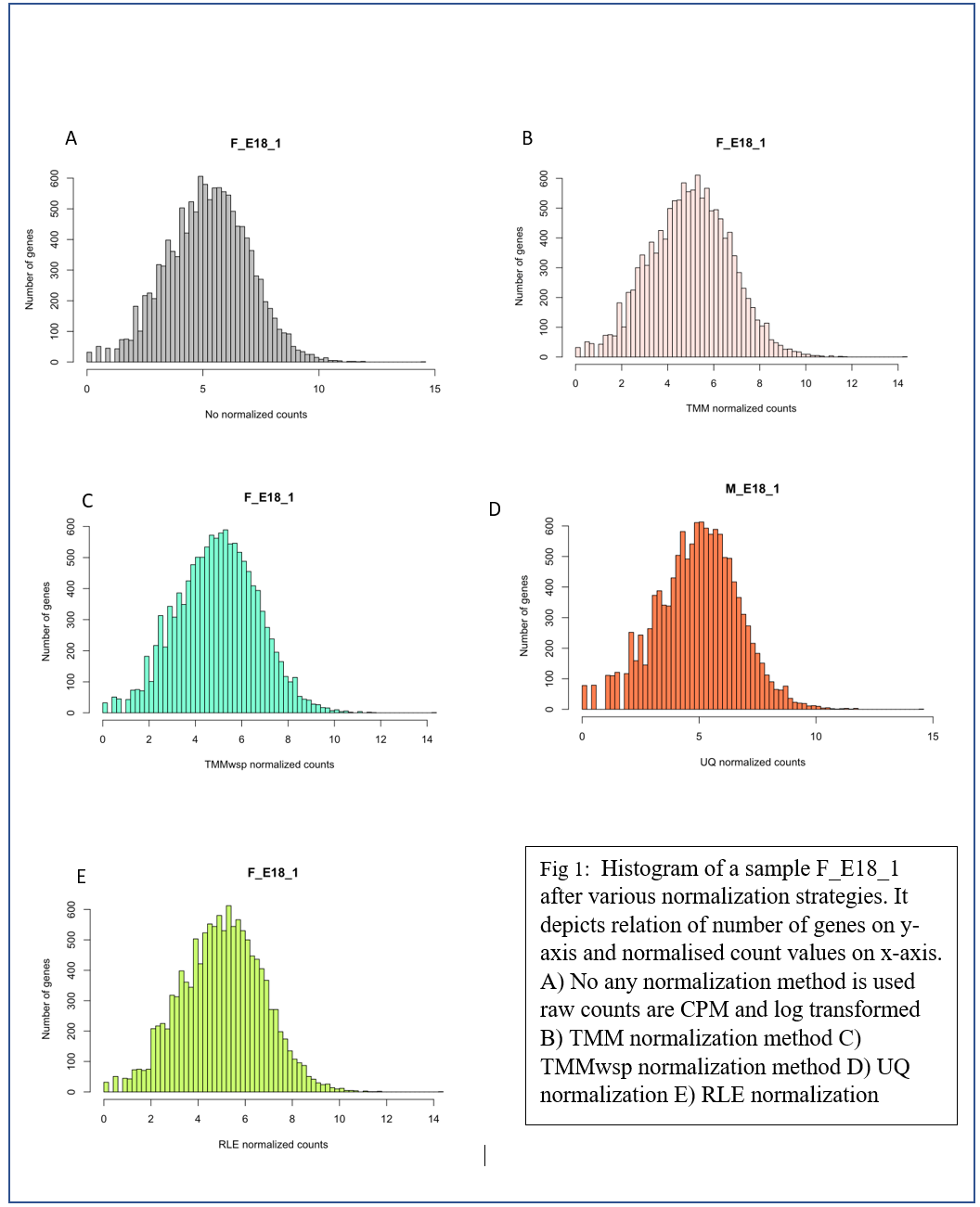
Likelihood of presence of batch effect on large dataset such as one being analysed here is high. ComBat seq in *SVAseq* package was used. ComBat seq uses empirical Bayes method to estimate parameters to adjust the batches and output are positive integers making easy to use for downstream analysis.

The MDS (Multidimensional scaling) plot was used to explore difference between the batch corrected and uncorrected data. MDS clusters the similar samples together.

# Results

1. **No significant differences in the distribution of the samples**

Significant differences in the distribution of the samples were not observed using different normalization techniques as an example the histogram for F\_E18\_1 (Female embryonic day 18 replicate 1) is shown. (Code with all the histograms of all the samples are in (code and analysis:[Distribution\_plots.html](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/Hamansagar%20Normalization%20Files/Distribution_plots.html?csf=1&web=1&e=CimpPL) for developmental phase and [Distribution\_plots\_treated\_samples.html](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/Hamansagar%20Normalization%20Files/Treated_samples/Distribution_plots_treated_samples.html?csf=1&web=1&e=3cKboF) for developmental phase)



1. **Only negative control data have fluctuation of median and quartile values**

All the samples had similar median values observed in the boxplots after all four normalisation approaches were implemented except for the boxplot of the negative control where slight fluctuation was observed. This showed that four approaches are making the samples more comparable than no normalization. The abline is created from median value of the entire dataset. (Code and analysis: [normalisation\_strategy.html](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/Hamansagar%20Normalization%20Files/normalisation_strategy.html?csf=1&web=1&e=0jPNdd) for developmental phases and [normalisation\_strategy\_treated\_samples.html](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/Hamansagar%20Normalization%20Files/Treated_samples/normalisation_strategy_treated_samples.html?csf=1&web=1&e=wKJD2V) for treated samples)

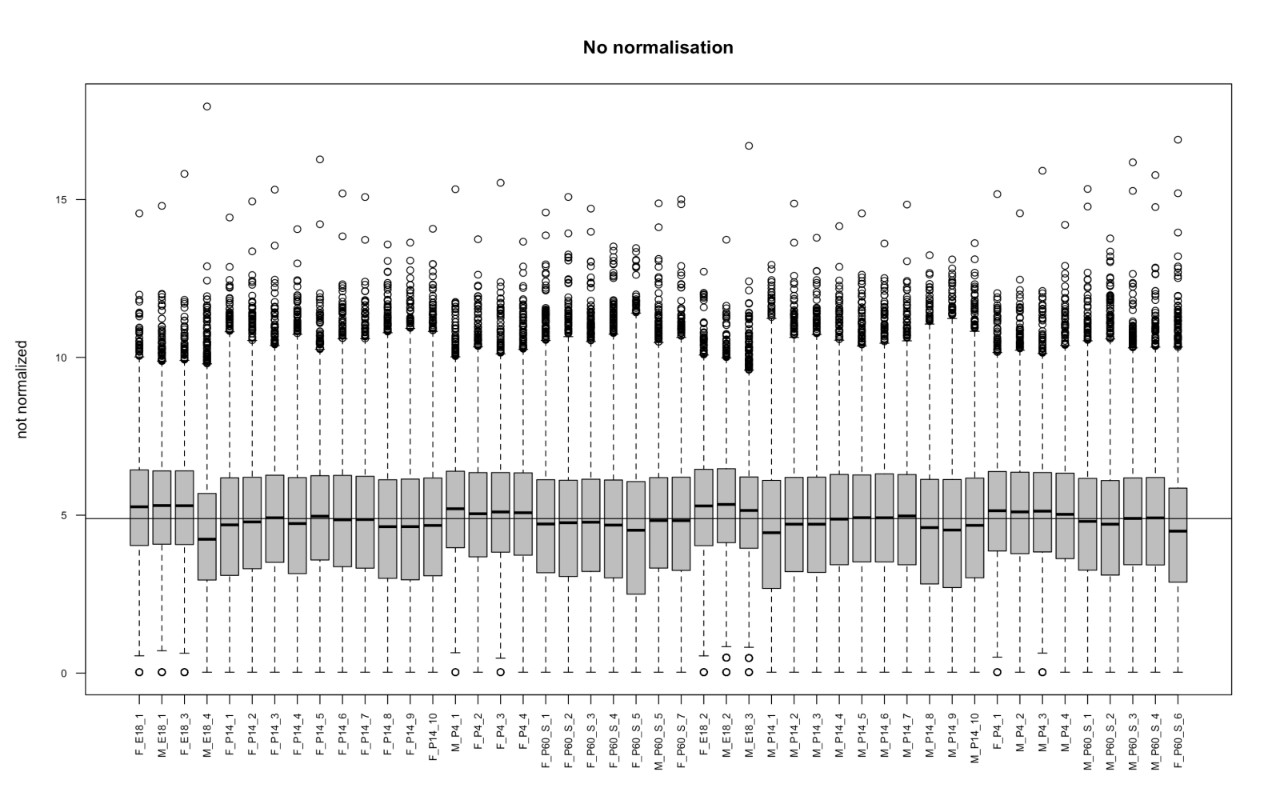


Fig 2: Box plot for the developmental phase showing the non-normalized counts on y axis and samples on x axis. Median values are slightly fluctuating from abline.

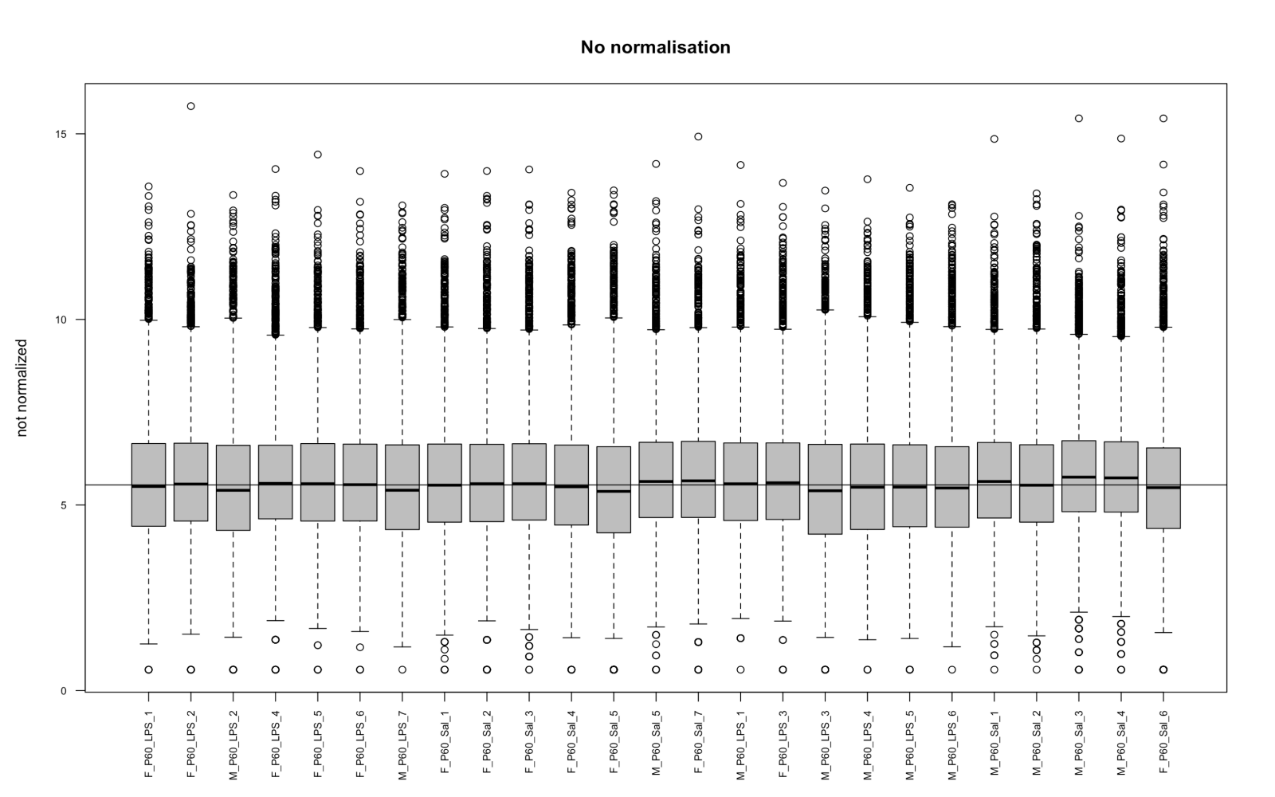


Fig3: Box for plot LPS and Saline treated cells showing the non-normalized counts on y axis and samples on x axis. Median values are slightly fluctuating from abline.

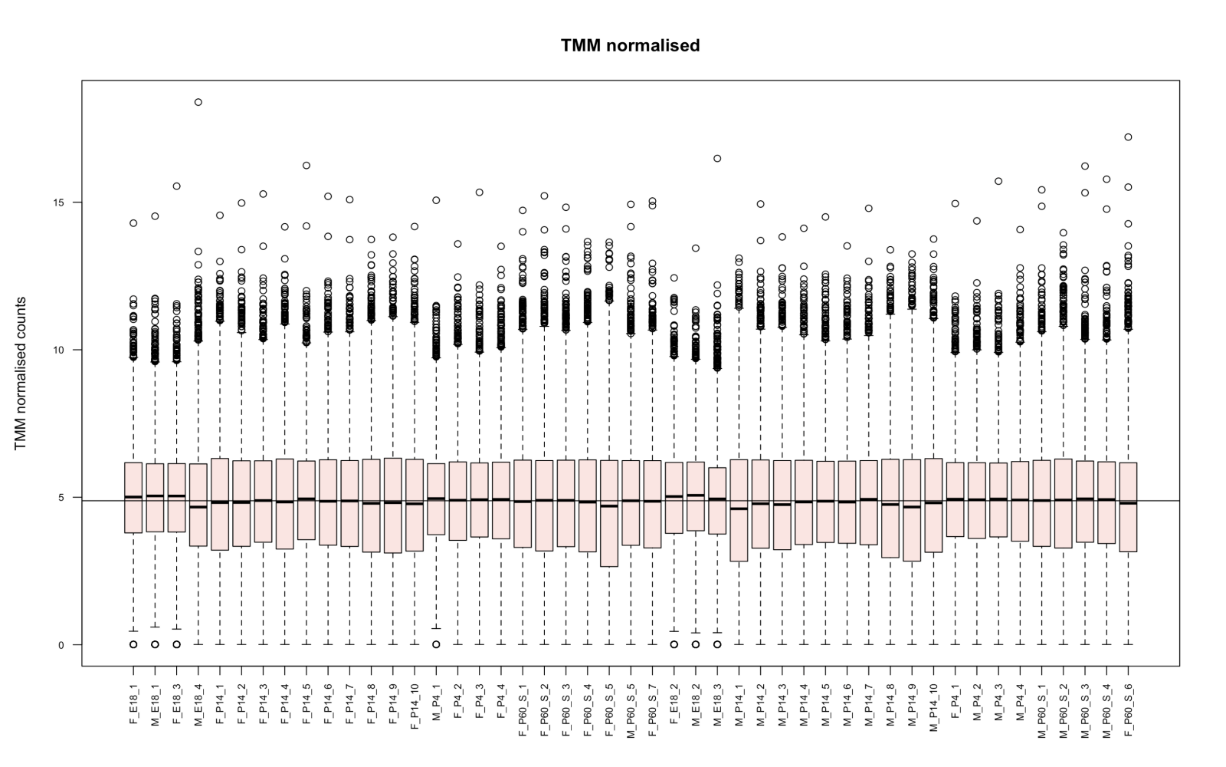


Fig 4: Box plot for the developmental phase showing the TMM normalized counts on y axis and samples on x axis. Median values are not so much fluctuating from abline.

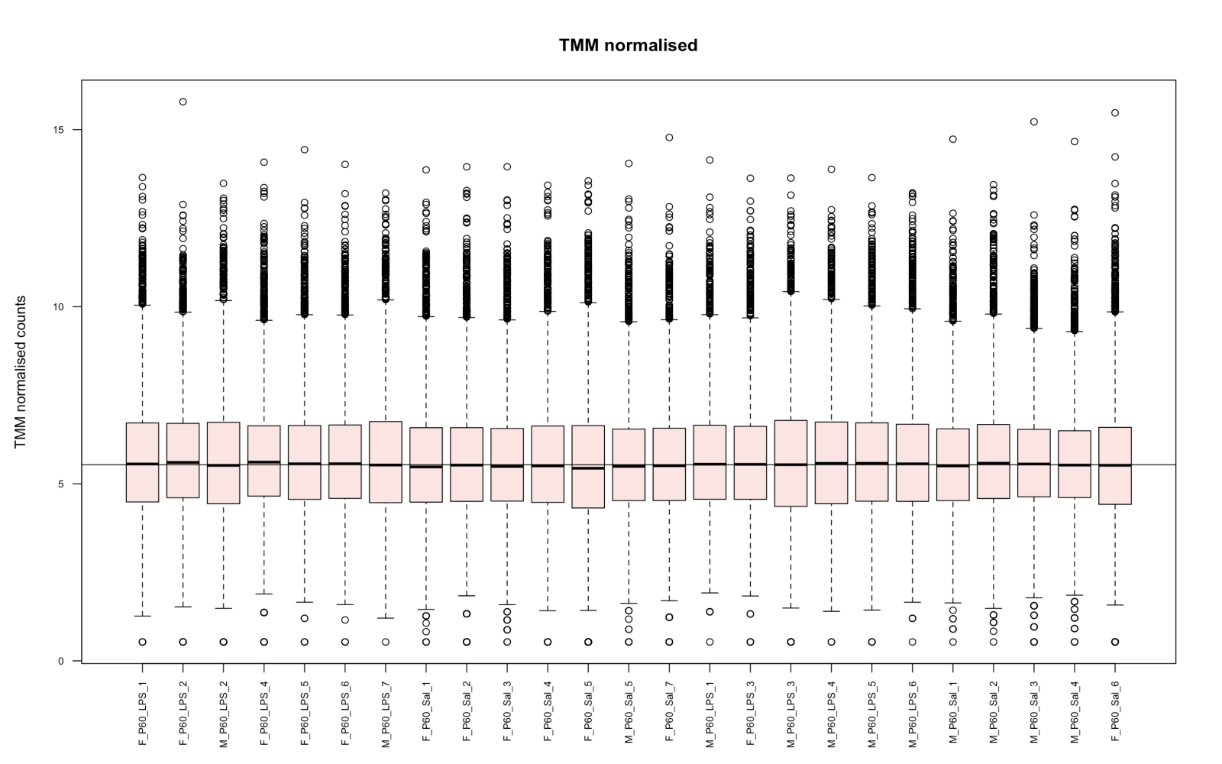


Fig 5: Box plot for LPS and Saline treated cells showing the TMM normalized counts on y axis and samples on x axis. Median values are not fluctuating from abline.

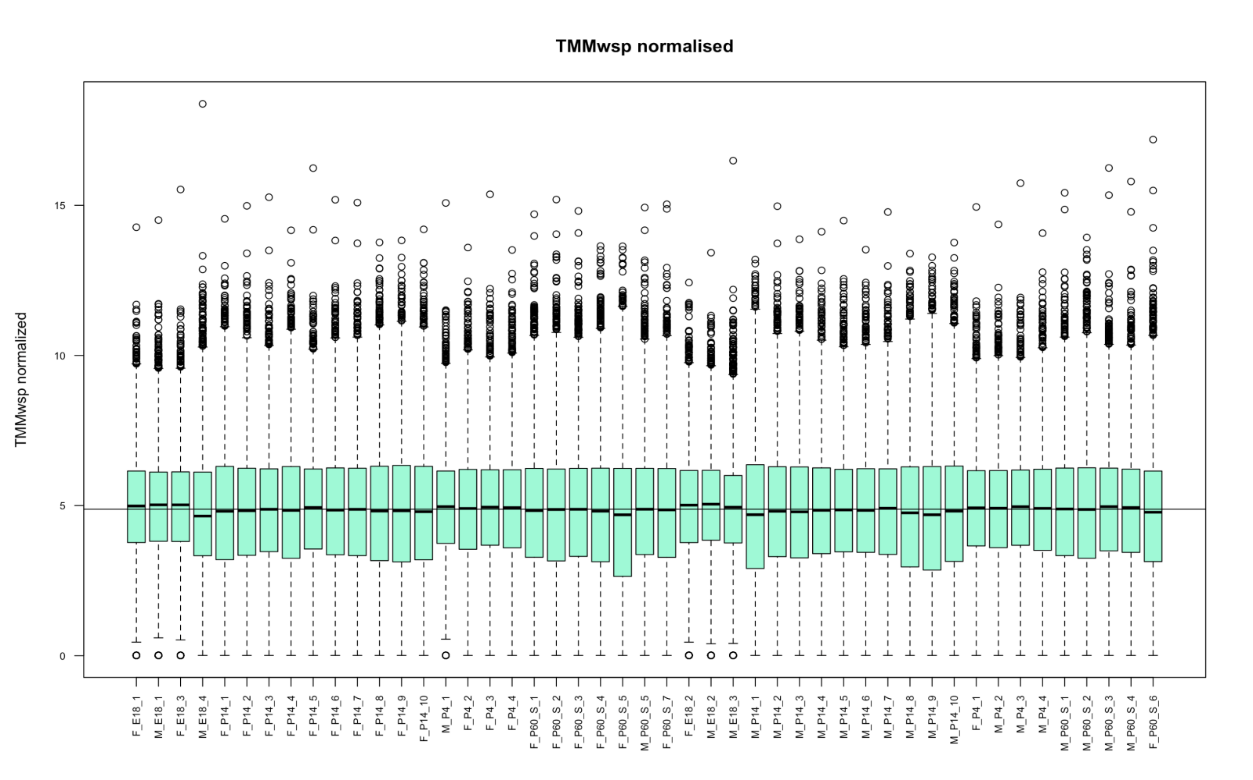


Fig 6: Box plot for the developmental phase showing the TMM normalized counts on y axis and samples on x axis. Median values are not so much fluctuating from abline.

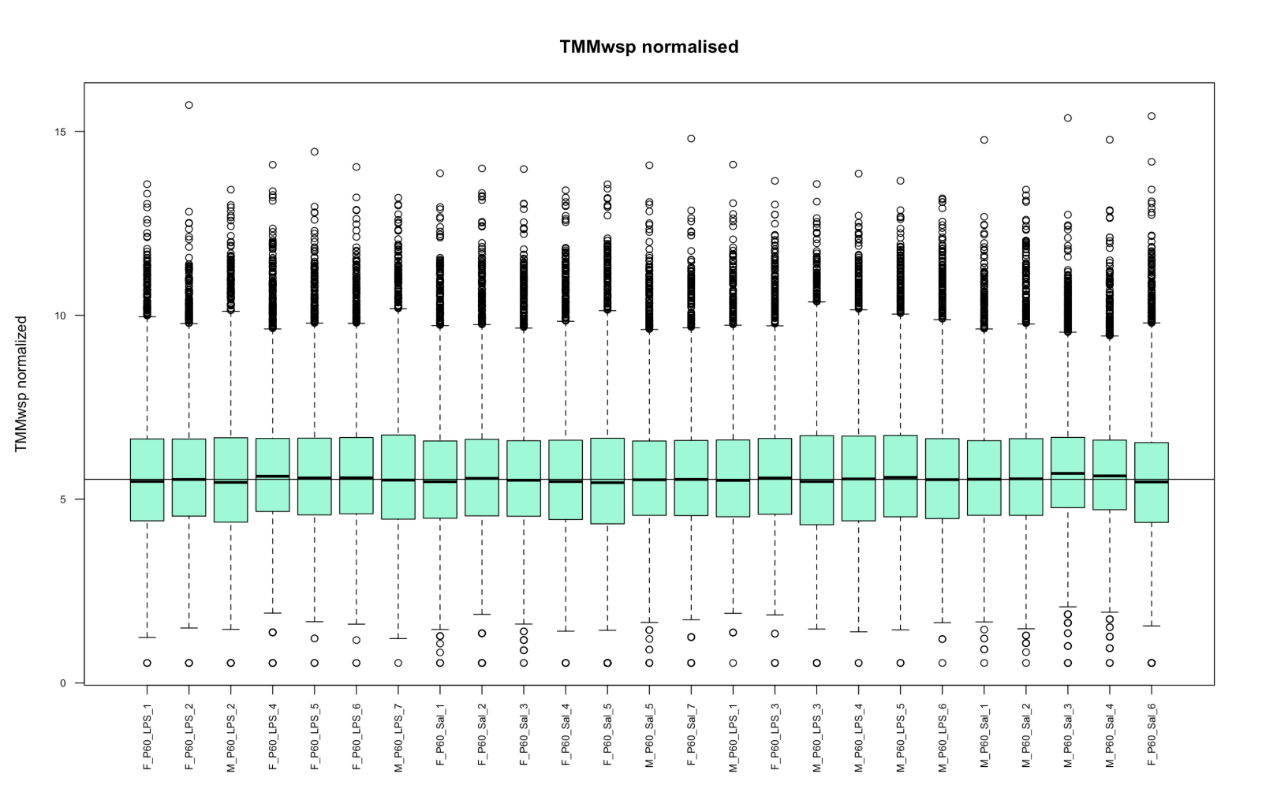


Fig 7: Box plot for LPS and Saline treated cells showing the TMMwsp normalized counts on y axis and samples on x axis. Median values are not fluctuating from abline.

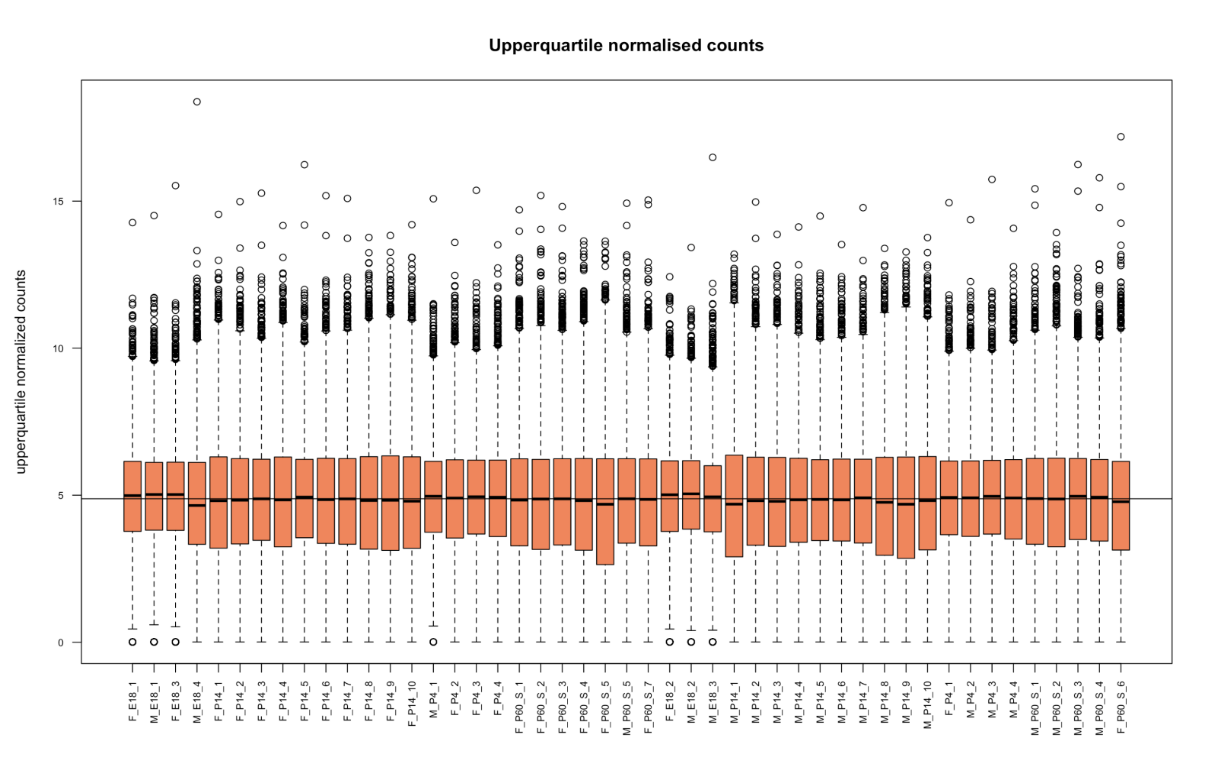


Fig 8: Box plot for the developmental phase showing the UQ normalized counts on y axis and samples on x axis. Median values are not fluctuating from abline.

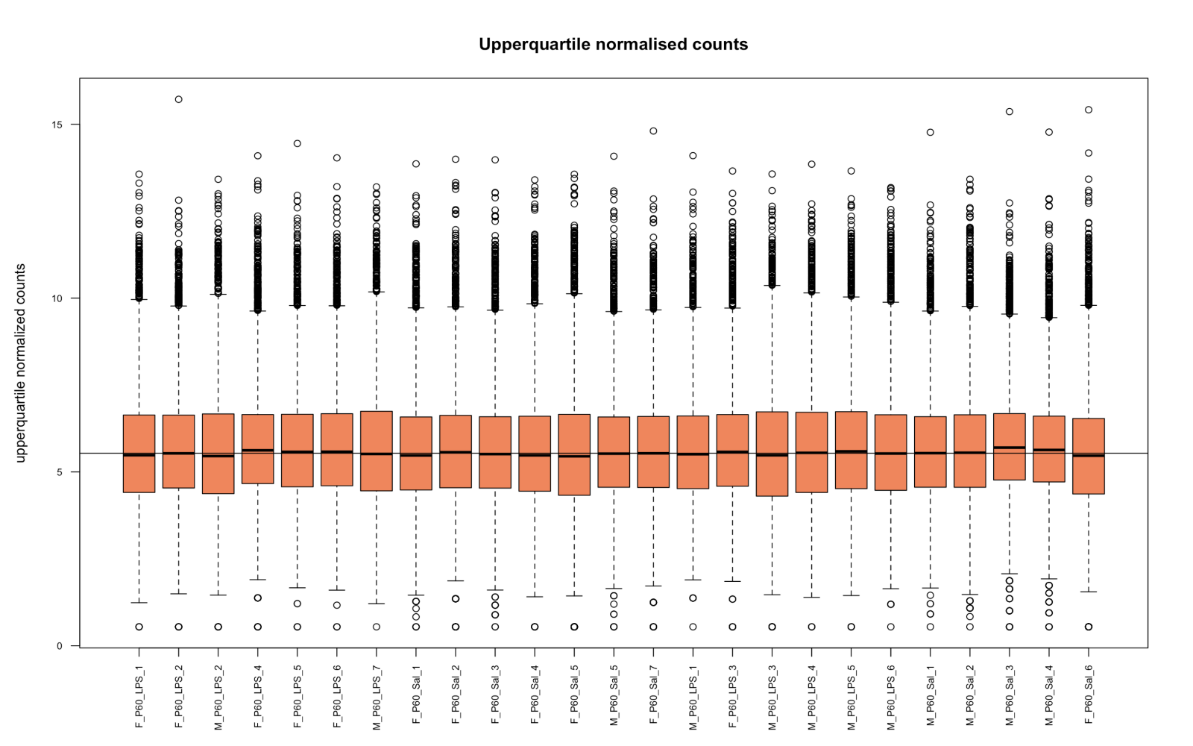


Fig 9: Box plot LPS and Saline treated cells showing the UQ normalized counts on y axis and samples on x axis. Median values are not so much fluctuating from abline.

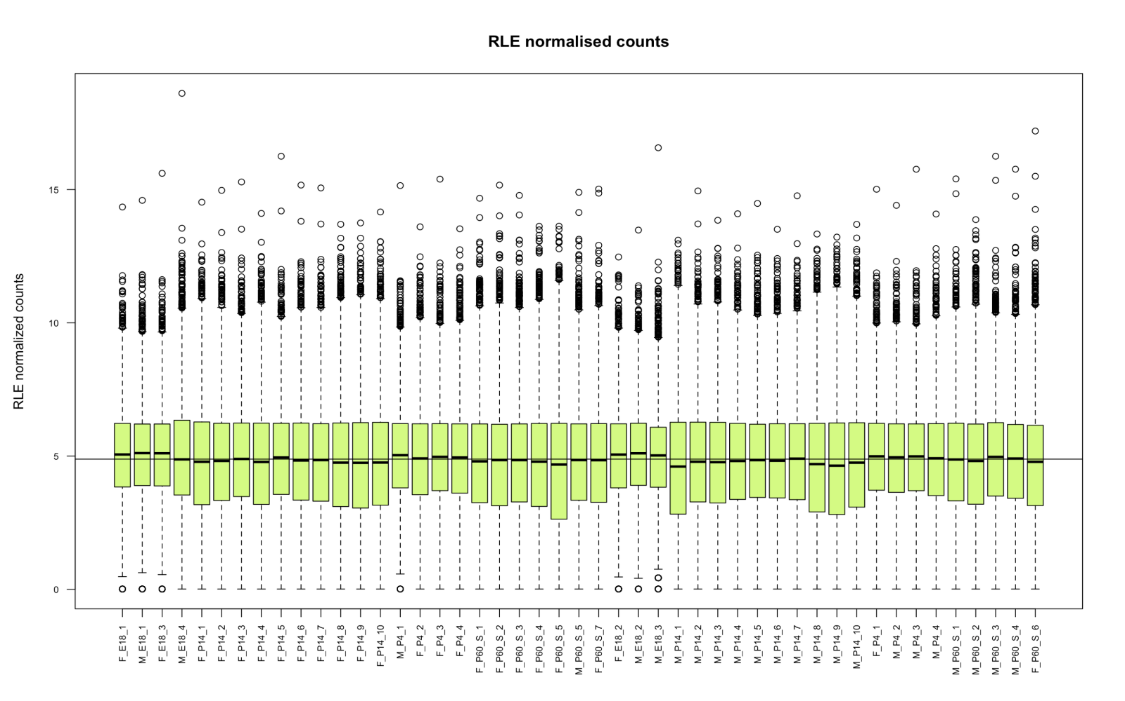


Fig 10: Box plot for the developmental phase showing the RLE normalized counts on y axis and samples on x axis Median values are not so much fluctuating from abline.

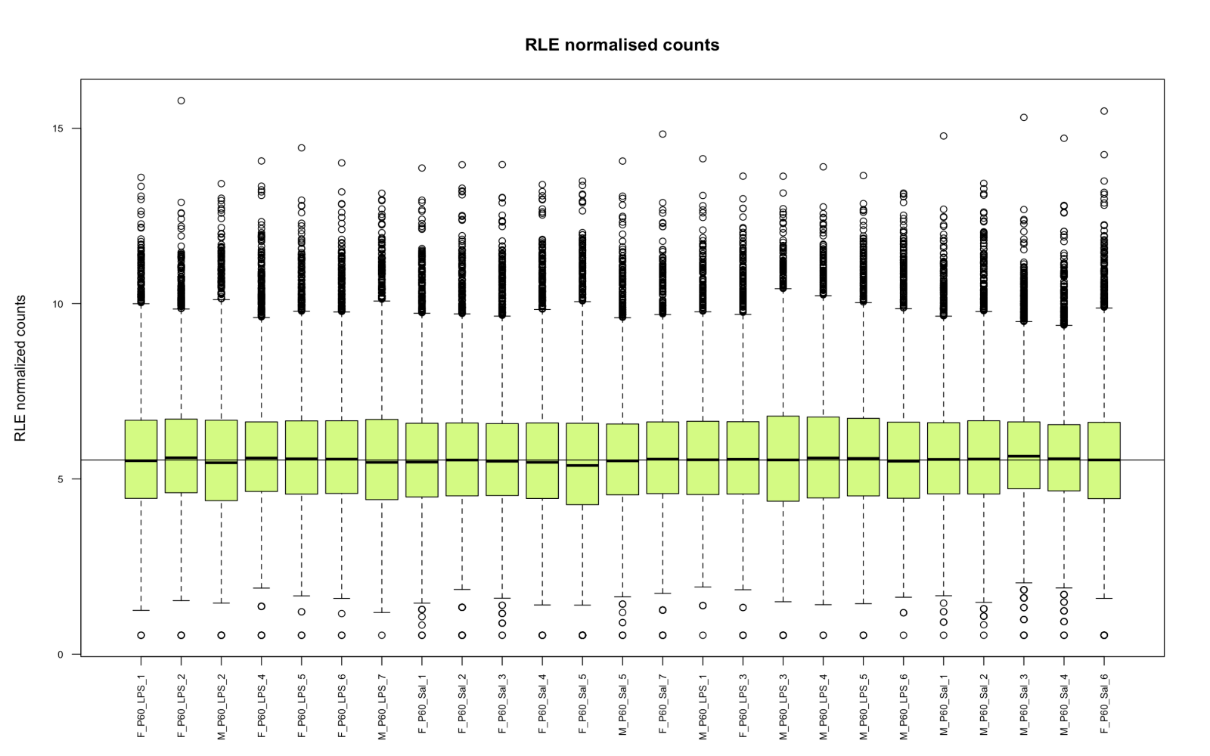


Fig 11: Box plot for LPS and Saline treated cells showing the RLE normalized counts on y axis and samples on x axis. Median values are not fluctuating from abline.

1. **Difference was observed in differential gene expression analysis**

For developmental phase, to get differentially expressed genes (DEGs) contrast matrix was set E18 versus P4. Highest number of DEGs, 6499, were obtained from RLE normalization, which was 10.63% higher than common genes between all the techniques, followed by UQ, TMMwsp and TMM. RLE and UQ are the techniques which does not trim the extreme values like TMM and TMMwsp does. The lowest number of DEGs were obtained from negative control, only 5978 total DEGs. Net 114 unique DEGs were obtained from TMMwsp which is predictable result as its algorithm includes the genes which have at least one positive count while comparing for pair of libraries, whereas TMM ignores gene if it encounters the zero count while comparing a pair of libraries. This also explains why the TMM has no unique set of genes since the dataset has inflated zeros. The detail of differential gene expression is in [normalisation\_strategy.html](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/Hamansagar%20Normalization%20Files/normalisation_strategy.html?csf=1&web=1&e=0jPNdd) and set of unique genes along with fold change and P values are in [Unique\_elements\_of\_set.html.](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/Hamansagar%20Normalization%20Files/Unique_elements_of_set.html?csf=1&web=1&e=qvA3jX)

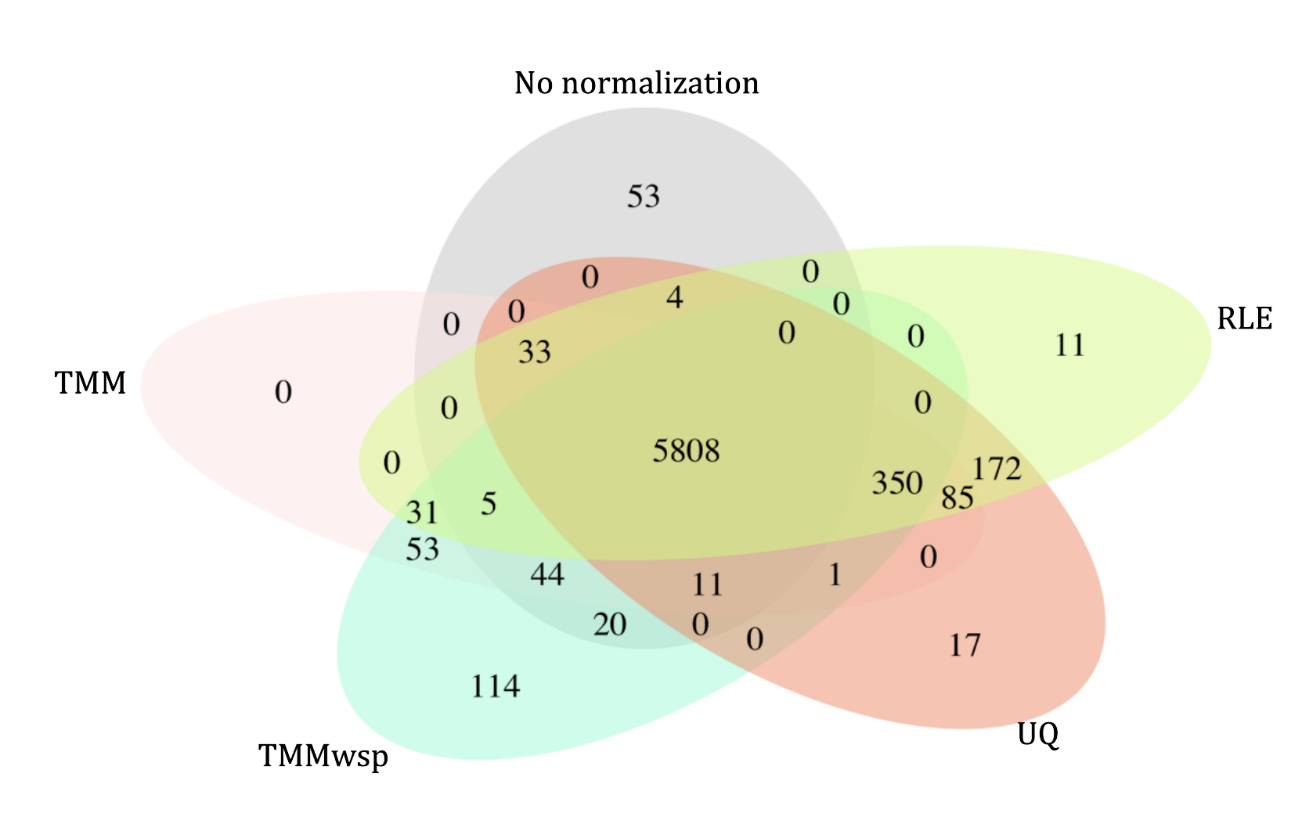


Fig 12: Venn diagram for the developmental phase of DEGs using various normalization strategies. 5808 genes are overlapping between all methods and control. Total number of DEGs detected are 5978,6421, 6437, 6481, 6499 respectively from no normalization, TMM, TMMwsp, UQ and RLE techniques.

For LPS and Saline treated samples, DEGs were obtained by contrast matrix set P60\_sal versus P60\_LPS. Highest number of DEGs, were obtained from negative control that is 1369 normalization, followed by UQ, RLE, TMM and TMMwsp. Code for analysis [normalisation\_strategy\_treated\_samples.html](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/Hamansagar%20Normalization%20Files/Treated_samples/normalisation_strategy_treated_samples.html?csf=1&web=1&e=teNCzm)[)](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/Hamansagar%20Normalization%20Files/Treated_samples/normalisation_strategy_treated_samples.html?csf=1&web=1&e=VKrpny)

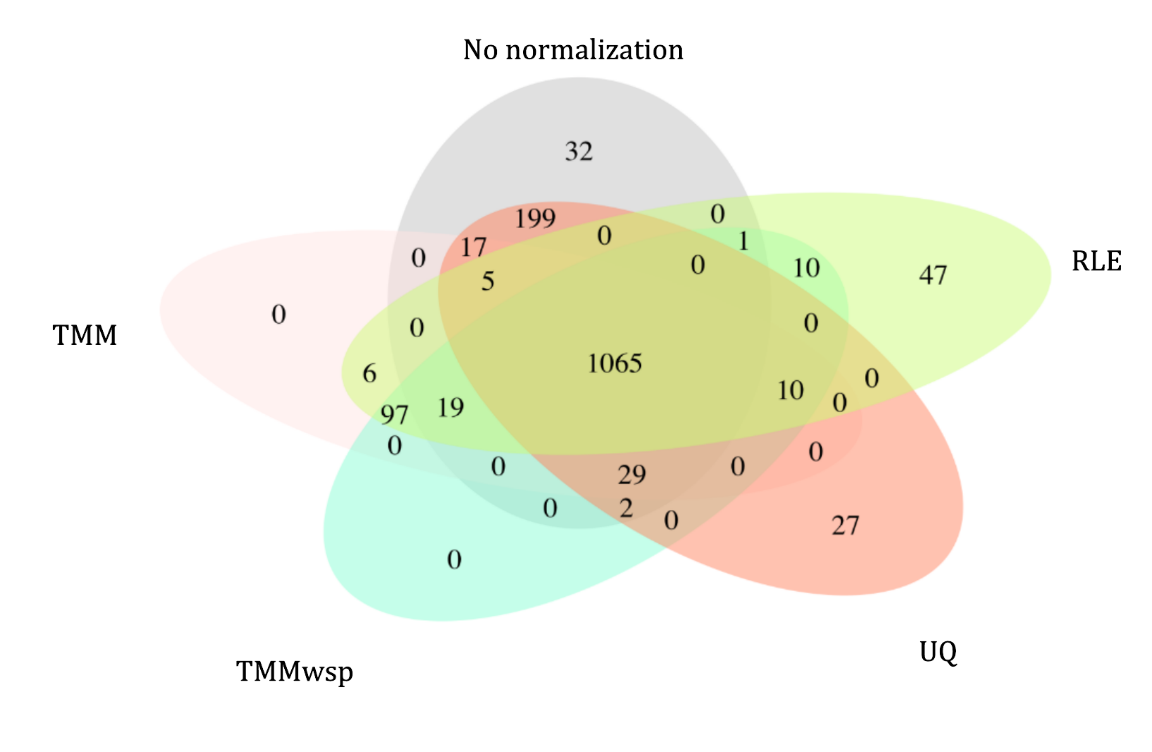


Fig 12: Venn diagram of for LPS and Saline treated cells DEGs using various normalization strategies. 1065 genes are overlapping between all methods and control. Total number of DEGs detected are 1369, 1354, 1260,1248 1233, respectively from no normalization, UQ, RLE, TMM, TMMwsp techniques.

1. **Batch effect correction with ComBat seq did not produce meaning result**

From MDS plot after correcting for the batch effect by ComBat seq, variance among the sample was significantly decreased, approximately by 21% and clustering of P4 samples with P14 samples was observed which is biologically unreasonable. Therefore, the differential gene expression was performed without batch correction of this dataset. Code for batch effect correction in [Combat\_seq\_new.html](https://rmiteduau-my.sharepoint.com/:u:/r/personal/s3971121_student_rmit_edu_au/Documents/CASP/Hamansagar%20Normalization%20Files/Combat_seq_new.html?csf=1&web=1&e=c33Uwi)

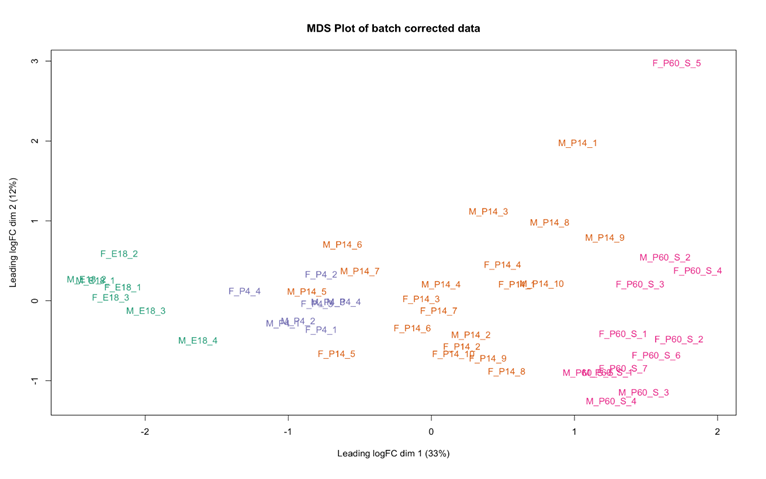


Fig 8: MDS plot of batch corrected data. The x axis is representing dimension 1 of batch corrected values which is CPM normalized and log transformed, and a y axis is presenting dimension 2. Colouring by age group. Variance in dimension 1 is 33% and P4 samples are clustered with P14 makes the process less reliable for this dataset.

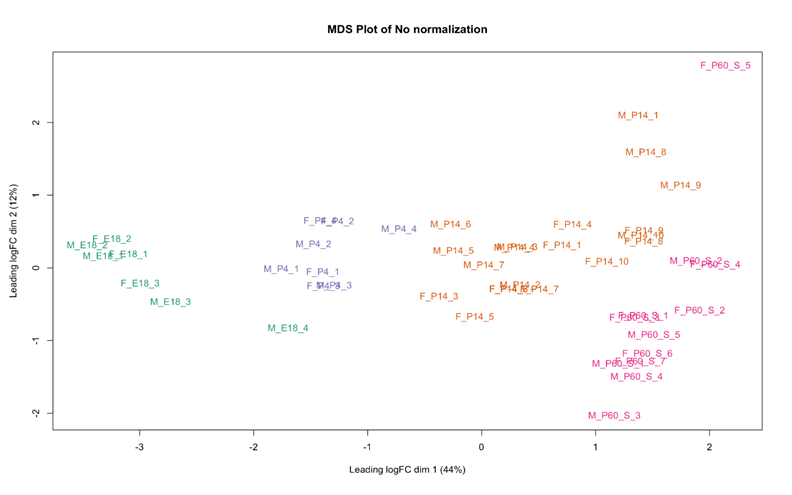


Fig 9: MDS plot of data without batch effect correction. The x axis is representing dimension 1 of values which is CPM normalized and log transformed, and a y axis is presenting dimension 2. Colouring by age group. Variance in dimension 1 is 44% and samples are clustered according to age on dim 1.

# Discussion

Normalization, is not a single step process, is rather a layer of multiple processes to remove unwanted technical variations until data is clean with biological variation alone. First the lowly expressed genes are filtered. The presence of lowly expressed genes is indistinguishable to noise which decreases the sensitivity of differentially expressed genes (DEGs) (McIntyre et al. 2011; Sha, Phan & Wang 2015). After correcting sequencing depth variation, compositional bias and other known and unknown technical variations, then, data can be used for differential gene expression analysis. When most genes are not differentially expressed then methods such as TMM, RLE and Quantile normalization can be used (Evans, Hardin & Stoebel 2017; Zhao, Yaxing, Wong & Goh 2020). These three methods are different on how the reference sample is chosen and how their scaling factor is calculated to the adjust outlier in the data. TMM trims extreme values, RLE uses geometric mean and median values to handle the outliers whereas quantile uses 75% quantile of each count.

From the results we have, most of the DEGs we got from different techniques were same. However, among them, the algorithm of TMM and TMMwsp is more robust for handling the outliers, as it trims the outlier data. If the data has inflated zeros, then its variant TMMwsp is more appropriate. It is crucial that a thorough evaluation of normalization techniques is required before conducting downstream statistical analysis. The best application of the technique depends on nature of data. If the data fulfils the statistical assumption, then they are good to use. In this case, all the four methods are appropriate for this data as fulfils the statistical assumption that majority of genes are not differentially expressed. However, the best way to verify, if the accuracy and precision of normalization method is by conducting RT-qPCR.

# Conclusion

Since, majority of differentially expressed genes are overlapping, any one of the four techniques can be used for this dataset. The best way to determine which technique is the most precise and accurate showing the true biological signal would be conducting the RT -qPCR for the unique gene sets we have from the Venn diagram. From the statistical point of view, TMMwsp has a more inclusive algorithm, it does not remove genes while comparing the two libraries even if one of them has zero count, unlike TMM which does remove. I think TMMwsp can be used as method of preference to minimize gene loss due to zeros for large scale studies. However, for validation of this *in silico* conclusion, *in vitro* or *in vivo* experiments must be conducted.

**Identification of differentially expressed genes**

Finding differentially expressed genes (DEGs) between conditions is crucial to understand the molecular basis of phenotype variation (Soneson & Delorenzi 2013). DEGs are identified from the results of P-adjusted and expression fold change values by using parametric tests and non-parametric tests. Parametric tests include the pairwise comparison between a gene from the two samples. DESeq2 (Love, Huber & Anders 2014), edgeR (Chen, Lun & Smyth 2016), and Limma-voom (Law et al. 2014) packages are based on pairwise parametric test. Another class of parametric test is ANOVA (Analysis of variance), ANOVA is used to assess the variation in gene expression across multiple groups. The null hypothesis for ANOVA states that the gene expression is uniform across all the groups, while the alternative hypothesis posits that there are significant differences in gene expression among the groups. Nonparametric methods are data adaptive which does not depend on distribution of the count data. Some packages for non-parametric tests are SAMseq (Li, J & Tibshirani 2011), NPEBseq (Bi & Davuluri 2013), NOIseq (Tarazona et al. 2015). These approaches employ distinct strategies: SAMseq utilizes the rank of expression values in a Wilcoxon statistic, NPEBseq adopts an empirical Bayesian approach based on expression fold change for the test statistic, and NOIseq uses fold change and difference in expression as a test statistic. Among all DESeq2, edgR and Limma voom are the most widely used. Although a recently published paper showed DESeq2 and edgeR produce drastically high number false positives in experiments with sample size more than 100 (Li, Y et al. 2022). However, this study has a major drawback because in large scale studies, unwanted batch effects and variations are highly likely to occur. This study also has not filtered lowly expressed genes while using DESeq2 pipeline, which might be a possible reason for a high number of false positives in its results. Also, non-parametric tests are not as powerful as parametric tests and are less sensitive.

Before, doing the any test statistic, the count data is modelled. The count data from RNA-seq is highly dispersed so the natural choices for a modelling the count data either Poisson or Binomial, due to its discrete nature. However, a negative binomial (NB) models are fitted for the count data as mean is greater than variance (Anders & Huber 2010; Lu, Tomfohr & Kepler 2005; Robinson & Smyth 2007). In case, if mean was equal to variance, Poisson distribution would be the best fit.  Both DESeq2 and edgeR are based on modelling on NB, employ similar methods and produce similar results, whereas Limma implements linear modelling. Hence, in this review, we will compare three popular methods based on parametric tests, they are DESeq2, edgeR and Limma voom.

1. **edgeR**

edgeR is based on negative binomial distribution and applies empirical Bayes methods, exact tests and generalized linear models (GLMs). It has a range of options for normalization of data such as, TMM, TMMwsp, RLE and UQ (cite). The classic technique includes quantile adjusted conditional likelihood method for simple experiments with single factor. In more complicated experiments, researchers use GLMs, which are extensions of the traditional linear models designed to handle data that don't follow a normal distribution. GLMs use probability distributions based on how the variance changes with the mean of the data, allowing them to fit log-linear models and better analyze count data. (McCarthy, Chen & Smyth 2012).  In RNA-seq data, dispersion estimates represent the variability or spread in gene expression corresponding to a specific mean value. That is dispersion for genes with same means vary only in variance. For estimating dispersions, edgeR uses the Cox-Reid (CR) profile adjusted likelihood method (McCarthy, Chen & Smyth 2012). The CR method addresses multiple factors by utilizing a GLM with a design matrix. By using this approach, CR enables accurate estimation of the dispersion, taking into consideration systematic sources of variation in count data. When the number of replicates is less, the CR method employs the empirical Bayes (EB) approach, which shares information between genes to enhance the estimation process. The function *estimateDisp()* with a specified design is used to calculate variation for common, trended and gene/tag wise dispersions. For multifactor experiment, *estimateGLMTagwiseDisp()*, is highly recommended by authors.  Once the dispersion estimates are obtained and negative binomial GLM is fitted, Quasi likelihood F-test or likelihood ratio test is performed. Among them Quasi likelihood F-test is preferred one as it is robust and error rate is controlled even if the number of replicates is less (Lun, Chen & Smyth 2016).

1. **DESeq2**

Similarly, DESeq2 also utilizes NB to model raw count data. In DESeq2, the count data is normalized by default RLE normalization.  It proceeds to estimate gene-wise dispersions, which are subsequently shrinked to yield more precise dispersion. Finally, conducting hypothesis through the Walt test or LRT (Likelihood Ratio Test) (Love, Huber & Anders 2014).   DESeq2, unlike edgeR, uses a method called “shrinkage” for dispersion estimation. This process involves sharing information among genes to obtain a more precise estimation of variation based on each gene's mean expression level. The dispersion is estimated using maximum likelihood estimation. This shrinkage method is crucial for reducing false positives in differential analysis. It involves shrinking the gene-wise dispersion estimates and fitting a GLM for each gene (Love, Huber & Anders 2014). Vst (Variance stabilizing transformation)and rlog (Regularized log), both stabilize the variance making the data more homoscedastic, having constant variance across the mean values. The rlog transformation is less affected by variations in scaling factors, which can be problematic when size factors have wide differences, hence the preference is given vst. DESeq2 provides two hypothesis testing methods for differential expression analysis, the Wald test and LRT (Likelihood ratio test). The Wald test is a test involving the division of the log-fold change by its corresponding standard error, thereby generating a z-statistic. The wald test is the default method, in DESeq2 unless LRT is mentioned. The second method is the LRT, a powerful tool used to assess multiple terms or factors concurrently within a statistical model. The LRT is especially advantageous when examining complex experimental designs involving several variables or conditions simultaneously.

1. **Limma voom**

Unlike the previously mentioned methods that employ GLM, Limma follows a different approach by using linear models to work with the count data (Law et al. 2014). Limma finds out the relationship between the mean and the variance by considering logarithmic counts, calculates precise weights for each observation, and integrates these weights into the Limma empirical Bayes analysis. A significant advantage of Limma is its capability to handle complicated experiments involving various treatment factors. Mean-variance relationship is derived non parametrically from normalized log counts either by Limma trend (modified Limma empirical Bayes process to incorporate the mean-variance relationship for gene level) or Limma voom (incorporates the mean-variance trend into a precision weight for each individual normalized observation). When working with data from different sequencing depths, the voom method is more effective than the trend method. This is because even if the normalized values appear similar, the actual counts can be quite different due to variations in sequencing depth. Therefore, when trying to understand how the average and variance of the data are related, it's better to use the voom method for each individual data point (Law et al. 2014). The estimated weight after voom transformation is used in the Limma empirical Bayes analysis pipeline. The gene expression differential analysis between groups is tested using the moderated t-statistics with robust as true to reduce to effect of outlier data (Belinda et al. 2016).

1. **Multiple test correction**

Since, there are thousands of genes in RNA seq data, multiple test correction is necessary to control the false positives or type I error. False Discovery rate (FDR)/Benjamini-Hochberg (BH) is widely used method to control number of false positives. In this method, genes are ranked from smallest to largest based on p values, then multiply each ranked p value by (total number of tests / rank) (Benjamini & Hochberg 1995). The null hypothesis for test statistic is there is no differential expression across the two samples. If adjusted p value / FDR is small, less than 0.05, null hypothesis is rejected which means the gene is differentially expressed. Hence, multiple test correction is important to know differential gene expression.

**Comparison of the three methods**

edgeR and DESeq2 are both based on NB GLM technique where Limma is based on Linear modelling strategy. All of three techniques uses empirical Bayes theorm to shrink the dispersion of overly dispersed count values. DESeq2 and Limma has a step where the variances are further stablised to account the over dispersion, which is not available in edgeR. All the three approaches use parametric test to conclude if the gene is differentially expressed or not. edgeR and DESeq2 yields similar results but DESeq2 stabilize the variance. Hence the recommended method for Richa’s data analysis is DESeq2.

# References

Aird, D, Ross, MG, Chen, W-S, Danielsson, M, Fennell, T, Russ, C, Jaffe, DB, Nusbaum, C & Gnirke, A 2011, 'Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries', *Genome Biology*, vol. 12, no. 2, pp. 1-14.

Benjamini, Y & Speed, TP 2012, 'Summarizing and correcting the GC content bias in high-throughput sequencing', *Nucleic acids research*, vol. 40, no. 10, pp. e72-e72.

Bolstad, BM, Irizarry, RA, Åstrand, M & Speed, TP 2003, 'A comparison of normalization methods for high density oligonucleotide array data based on variance and bias', *Bioinformatics*, vol. 19, no. 2, pp. 185-193.

Bullard, JH, Purdom, E, Hansen, KD & Dudoit, S 2010, 'Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments', *BMC Bioinformatics*, vol. 11, no. 1, p. 94.

Chen, Y, Lun, A & Smyth, G 2016, 'From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; peer review: 5 approved]', *F1000Research*, vol. 5, no. 1438.

Conesa, A, Madrigal, P, Tarazona, S, Gomez-Cabrero, D, Cervera, A, McPherson, A, Szcześniak, MW, Gaffney, DJ, Elo, LL, Zhang, X & Mortazavi, A 2016, 'A survey of best practices for RNA-seq data analysis', *Genome Biology*, vol. 17, no. 1, p. 13.

Dillies, M-A, Rau, A, Aubert, J, Hennequet-Antier, C, Jeanmougin, M, Servant, N, Keime, C, Marot, G, Castel, D, Estelle, J, Guernec, G, Jagla, B, Jouneau, L, Laloë, D, Le Gall, C, Schaëffer, B, Le Crom, S, Guedj, M, Jaffrézic, F & on behalf of The French StatOmique, C 2013, 'A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis', *Briefings in Bioinformatics*, vol. 14, no. 6, pp. 671-683.

Evans, C, Hardin, J & Stoebel, DM 2017, 'Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions', *Briefings in Bioinformatics*, vol. 19, no. 5, pp. 776-792.

Grün, D & van Oudenaarden, A 2015, 'Design and Analysis of Single-Cell Sequencing Experiments', *Cell*, vol. 163, no. 4, pp. 799-810.

Hicks, SC, Okrah, K, Paulson, JN, Quackenbush, J, Irizarry, RA & Bravo, HC 2017, 'Smooth quantile normalization', *Biostatistics*, vol. 19, no. 2, pp. 185-198.

Law, C, Alhamdoosh, M, Su, S, Dong, X, Tian, L, Smyth, G & Ritchie, M 2018, 'RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR [version 3; peer review: 3 approved]', *F1000Research*, vol. 5, no. 1408.

Leek, JT 2014, 'svaseq: removing batch effects and other unwanted noise from sequencing data', *Nucleic acids research*, vol. 42, no. 21, pp. e161-e161.

Leek, JT, Scharpf, RB, Bravo, HC, Simcha, D, Langmead, B, Johnson, WE, Geman, D, Baggerly, K & Irizarry, RA 2010, 'Tackling the widespread and critical impact of batch effects in high-throughput data', *Nature Reviews Genetics*, vol. 11, no. 10, pp. 733-739.

Li, P, Piao, Y, Shon, HS & Ryu, KH 2015, 'Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data', *BMC Bioinformatics*, vol. 16, no. 1, p. 347.

Lin, Y, Golovnina, K, Chen, Z-X, Lee, HN, Negron, YLS, Sultana, H, Oliver, B & Harbison, ST 2016, 'Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster', *BMC Genomics*, vol. 17, no. 1, p. 28.

Liu, X, Li, N, Liu, S, Wang, J, Zhang, N, Zheng, X, Leung, K-S & Cheng, L 2019, 'Normalization methods for the analysis of unbalanced transcriptome data: a review', *Frontiers in bioengineering and biotechnology*, vol. 7, p. 358.

Love, MI, Hogenesch, JB & Irizarry, RA 2016, 'Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation', *Nature Biotechnology*, vol. 34, no. 12, pp. 1287-1291.

Love, MI, Huber, W & Anders, S 2014, 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, vol. 15, no. 12, p. 550.

Lovén, J, Orlando, David A, Sigova, Alla A, Lin, Charles Y, Rahl, Peter B, Burge, Christopher B, Levens, David L, Lee, Tong I & Young, Richard A 2012, 'Revisiting Global Gene Expression Analysis', *Cell*, vol. 151, no. 3, pp. 476-482.

Maza, E, Frasse, P, Senin, P, Bouzayen, M & Zouine, M 2013, 'Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: a matter of relative size of studied transcriptomes', *Communicative & integrative biology*, vol. 6, no. 6, p. e25849.

Mortazavi, A, Williams, BA, McCue, K, Schaeffer, L & Wold, B 2008, 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nature Methods*, vol. 5, no. 7, pp. 621-628.

Oshlack, A & Wakefield, MJ 2009, 'Transcript length bias in RNA-seq data confounds systems biology', *Biology Direct*, vol. 4, no. 1, p. 14.

Phipson, B, Zappia, L & Oshlack, A 2017, 'Gene length and detection bias in single cell RNA sequencing protocols [version 1; peer review: 4 approved]', *F1000Research*, vol. 6, no. 595.

Risso, D, Ngai, J, Speed, TP & Dudoit, S 2014, 'Normalization of RNA-seq data using factor analysis of control genes or samples', *Nature Biotechnology*, vol. 32, no. 9, pp. 896-902.

Risso, D, Schwartz, K, Sherlock, G & Dudoit, S 2011, 'GC-Content Normalization for RNA-Seq Data', *BMC Bioinformatics*, vol. 12, no. 1, p. 480.

Robinson, MD & Oshlack, A 2010, 'A scaling normalization method for differential expression analysis of RNA-seq data', *Genome Biology*, vol. 11, no. 3, p. R25.

Trapnell, C, Williams, BA, Pertea, G, Mortazavi, A, Kwan, G, van Baren, MJ, Salzberg, SL, Wold, BJ & Pachter, L 2010, 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nature Biotechnology*, vol. 28, no. 5, pp. 511-515.

Wagner, GP, Kin, K & Lynch, VJ 2012, 'Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples', *Theory in Biosciences*, vol. 131, no. 4, pp. 281-285.

Wang, D, Cheng, L, Wang, M, Wu, R, Li, P, Li, B, Zhang, Y, Gu, Y, Zhao, W, Wang, C & Guo, Z 2011, 'Extensive increase of microarray signals in cancers calls for novel normalization assumptions', *Computational Biology and Chemistry*, vol. 35, no. 3, pp. 126-130.

Zhao, S, Ye, Z & Stanton, R 2020, 'Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols', *Rna*, vol. 26, no. 8, pp. 903-909.

Zhao, Y, Li, M-C, Konaté, MM, Chen, L, Das, B, Karlovich, C, Williams, PM, Evrard, YA, Doroshow, JH & McShane, LM 2021, 'TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository', *Journal of Translational Medicine*, vol. 19, no. 1, p. 269.

Zhao, Y, Wong, L & Goh, WWB 2020, 'How to do quantile normalization correctly for gene expression data analyses', *Scientific Reports*, vol. 10, no. 1, p. 15534.

Zwiener, I, Frisch, B & Binder, H 2014, 'Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures', *PLOS ONE*, vol. 9, no. 1, p. e85150.

Zyprych-Walczak, J, Szabelska, A, Handschuh, L, Górczak, K, Klamecka, K, Figlerowicz, M & Siatkowski, I 2015, 'The impact of normalization methods on RNA-Seq data analysis', *BioMed research international*, vol. 2015.

Anders, S & Huber, W 2010, 'Differential expression analysis for sequence count data', *Nature Precedings*, vol.

￼

Belinda, P, Stanley, L, Ian, JM, Warren, SA & Gordon, KS 2016, 'Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression', *The Annals of Applied Statistics*, vol. 10, no. 2, pp. 946-963.

￼

Benjamini, Y & Hochberg, Y 1995, 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289-300.

￼

Bi, Y & Davuluri, RV 2013, 'NPEBseq: nonparametric empirical bayesian-based procedure for differential expression analysis of RNA-seq data', *BMC Bioinformatics*, vol. 14, no. 1, p. 262.

￼

Chen, Y, Lun, A & Smyth, G 2016, 'From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; peer review: 5 approved]', *F1000Research*, vol. 5, no. 1438.

￼

Law, CW, Chen, Y, Shi, W & Smyth, GK 2014, 'voom: precision weights unlock linear model analysis tools for RNA-seq read counts', *Genome Biology*, vol. 15, no. 2, p. R29.

￼

Li, J & Tibshirani, R 2011, 'Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data', *Statistical Methods in Medical Research*, vol. 22, no. 5, pp. 519-536.

￼

Li, Y, Ge, X, Peng, F, Li, W & Li, JJ 2022, 'Exaggerated false positives by popular differential expression methods when analyzing human population samples', *Genome Biology*, vol. 23, no. 1, p. 79.

￼

Love, MI, Huber, W & Anders, S 2014, 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, vol. 15, no. 12, p. 550.

￼

Lu, J, Tomfohr, JK & Kepler, TB 2005, 'Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach', *BMC Bioinformatics*, vol. 6, no. 1, p. 165.

￼

Lun, ATL, Chen, Y & Smyth, GK 2016, 'It’s DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR', in E Mathé & S Davis (eds), *Statistical Genomics: Methods and Protocols*, Springer New York, New York, NY, <<https://doi.org/10.1007/978-1-4939-3578-9_19>>, pp. 391-416.

￼

McCarthy, DJ, Chen, Y & Smyth, GK 2012, 'Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation', *Nucleic acids research*, vol. 40, no. 10, pp. 4288-4297.

￼

Robinson, MD & Smyth, GK 2007, 'Moderated statistical tests for assessing differences in tag abundance', *Bioinformatics*, vol. 23, no. 21, pp. 2881-2887.

￼

Soneson, C & Delorenzi, M 2013, 'A comparison of methods for differential expression analysis of RNA-seq data', *BMC Bioinformatics*, vol. 14, no. 1, p. 91.

￼

Tarazona, S, Furió-Tarí, P, Turrà, D, Pietro, AD, Nueda, MJ, Ferrer, A & Conesa, A 2015, 'Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package', *Nucleic acids research*, vol. 43, no. 21, pp. e140-e140.

￼

Law, CW, Chen, Y, Shi, W & Smyth, GK 2014, 'voom: precision weights unlock linear model analysis tools for RNA-seq read counts', *Genome Biology*, vol. 15, no. 2, p. R29.

Love, MI, Huber, W & Anders, S 2014, 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, vol. 15, no. 12, p. 550.